Heart Disease Prediction using

Machine Learning

*By*

Musale Krushna Pavan (Reg. No- 0000150)

Sanath Singavarapu (Reg. No- 0000188)

**Supervisor**: Dr. Bhaskar Biswas



*Bachelor Thesis submitted to*

Indian Institute of Information Technology Kalyani

*for the partial fulfillment of the degree of*

**Bachelor of Technology**

**in**

**Computer Science and Engineering**

**December 2019**

**Certificate**

This is to certify that the thesis entitled “Heart Disease Prediction using Machine Learning” being submitted by Sanath Singavarapu and Musale Krushna Pavan, an undergraduate student, Reg. No: 188 and 150, in the Department of Computer Science and Information Engineering, Indian Institute of Information Technology Kalyani, West Bengal 741235, India, for the award of Bachelor of Technology in Computer Science and Information Engineering is an original research work carried by him under my supervision and guidance. The thesis has fulfilled all the requirements as per the regulations of Indian Institute of Information Technology Kalyani and in my opinion, has reached the standards needed for submission. The work, techniques and the results presented have not been submitted to any other University or Institute for the award of any other degree or diploma

Dr. Bhaskar Biswas,

Faculty

Indian Institute Of Information Technology Kalyani

Declaration

I hereby declare that the work being presented in this thesis entitled, “Heart Disease Prediction using Machine Learning”, submitted to Indian Institute of Information Technology Kalyani in partial fulfillment for the award of the degree of **Bachelor of Technology** in Computer Science and Engineering during the period from June, 2019 to Dec, 2019 under the supervision of Dr. Bhaskar Biswas, Department of Computer Science and Engineering, Indian Institute of Information Technology Kalyani, West Bengal 741235, India, does not contain any classified information.

Name of the Candidate: Sanath Singavarapu, Musale Krushna Pavan.

Reg No/Roll No:188,150.

Name of the Department: Computer Science and Engineering.

Institute Name: Indian Institute of Information Technology Kalyani

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

..................................

Dr. Bhaskar Biswas,

Faculty

Indian Institute of Information Technology Kalyani

Place: Kalyani

Date:

Acknowledgments

First of all, I would like to take this opportunity to thank and to express our gratitude to our supervisor Dr. Bhaskar Biswas for his guidance, and instruction that she has given us from the time we have been doing thesis under his supervision through the completion of our undergraduate project. The skills we have able to learn from her during our Thesis work have benefited us immensely and will continue to do so throughout our future endeavors.

Name of the Candidate: Sanath Singavarapu, Musale Krushna Pavan.

Reg No/Roll No: 188, 150.

Name of the Department: Computer Science and Engineering.

Institute Name: Indian Institute of Information Technology Kalyani

Place: Kalyani.

Date:

**Abstract**

Heart disease is one of the most critical human diseases in the world and affects human life very badly. In today’s era deaths due to heart disease is a major issue, according to the survey approximately one person dies per minute due to heart disease. These statistics are considering all gendered category and also considering age group 25-70 these statistics may vary depending upon to the region and many other factors. [1]

Predicting the cause of the disease and occurrence of disease is a major challenge nowadays. Generally, Heart Disease Diagnosis require various medical tests and time consuming and involves our physical presence. As we are in the age of information and Machine learning and also computer field had got highly advanced. One common usage of machine learning is the prediction of an outcome based upon existing data. Mostly the machine algorithms learn patterns from the existing dataset, and then applies them to an unknown dataset in order to predict the outcome. [2]

Now we can use the tools and algorithms of machine learning and the considering the records of previous patients we can predict the occurrence of heart disease. Here in this paper, we have discussed various algorithms and tools used for prediction of heart diseases.

**Table of Contents**

1. Introduction
2. Machine Learning in Medical Field
   1. Supervised Learning
   2. Unsupervised Learning
3. Dataset Description
4. Algorithms Implemented
5. K Neighbors Classifier
6. Support Vector Machine
7. Decision Tree Classifier
8. Random Forest Classifier
9. Voting Classifier
10. Results
11. Future Scope
12. References

**Introduction**

The contents of this paper mainly focus on various datamining practices that are valuable in heart disease forecast with the assistance of dissimilar data mining tools that are accessible. If the heart doesn’t function properly, this will distress the other parts of the human body such as brain, kidney etc. Heart disease is a kind of disease which effects the functioning of the heart. In today’s era heart disease is the primary reason for deaths. WHO-World Health Organization has anticipated that 12 million people die every year because of heart diseases. Some heart diseases are cardiovascular, heart attack, coronary and knock. Knock is a sort of heart disease that occurs due to strengthening. blocking or lessening of blood vessels which drive through the brain or it can also be initiated by high blood pressure. [1]

Heart Disease is one of the most serious diseases that need to be taken in consideration and get aware to take necessary actions. Heart Disease is a life threat disease and predicting it in advance can save a human’s life. Heart Disease Check-up is expensive and time consuming as the doctors need to go through no of individuals tests records. In order to provide a solution and make it available to all we chose Heart Disease Prediction. Prediction of Heart Disease is a difficult task as there are many reasons which can lead to heart diseases. Out of the many reasons we considered some important factors for our project.

Some include: Blood Pressure, Cholesterol, Obesity etc.

Health Care today generates lot of data that can be used in combination with computer knowledge and produce some extremely good results.

**Machine Learning in Medical Field**

“Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.”

Machine Learning is mainly classified into two different groups:

**Supervised Learning:**

“Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labelled training data consisting of a set of training examples.”

For every given input X there is a labelled output Y. The model learns according to data by the programmed algorithm. The trained model predicts the output with an input given.

In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyses the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way (see inductive bias).

Given a set of ‘N’ training examples of the form ‘{x1, y1), ..., (xn,yn)}’ such that ‘x{i}’ is the feature vector of the i-th example and ‘ y{i}’ is its label (i.e., class), a learning algorithm seeks a function ‘g:X -> Y’ , where ‘X’ is the input space and ‘Y’ is the output space. The function ‘g’ is an element of some space of possible functions ‘G’, usually called the hypothesis space. It is sometimes convenient to represent ‘g’ using a scoring function ‘f: X x Y ->R’ such that ‘g’ is defined as returning the ‘y’ value that gives the highest score: ‘g(x)=arg max f(X,Y)’ . Let ‘F’ denote the space of scoring functions. [3]

Ex: K-Neighbours Classifier, Linear SVC etc.

**Unsupervised Learning:**

“Unsupervised learning is the training of an artificial intelligence (AI) algorithm using information that is neither classified nor labelled and allowing the algorithm to act on that information without guidance.”

The model learns features according to the given data. For every given input X there is no labelled output available.

This is a type of self-organized Hebbian learning that helps find previously unknown patterns in data set without pre-existing labels. It is also known as self-organization and allows modelling probability densities of given inputs.

Two of the main methods used in unsupervised learning are principal component and cluster analysis. Cluster analysis is used in unsupervised learning to group, or segment, datasets with shared attributes in order to extrapolate algorithmic relationships. Cluster analysis is a branch of machine learning that groups the data that has not been labelled, classified or categorized. Instead of responding to feedback, cluster analysis identifies commonalities in the data and reacts based on the presence or absence of such commonalities in each new piece of data. This approach helps detect anomalous data points that do not fit into either group. [4]

Ex: K-Means Clustering, Mixture Models, DBSCAN etc.

**Dataset Description**

Out of many factors that cause heart diseases we are considering some important factors that are mainly responsible. The dataset consists of total 14 parameters out of which 13 are input parameters and 1 is target value.

1. Age (In Years) Age of patient in years. 1: >=30 & =70

2. Sex 1: male, 0: female.

3. CP (Chest Pain) chest pain type Value 1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptomatic

4. Trestbps (resting blood pressure) Resting blood pressure (in mm/ Hg on admission to the hospital).

5. Chol (Cholesterol) Serum cholesterol in mg/dl (vh: very high; h: high)

6. Fbs (Fasting Blood Sugar) Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)

7. Restecg (Resting ECG result) Resting electrocardiographic results. Value 0: normal. Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV). Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria.

8. Thalach (Maximum heart rate) Maximum Heart Rate achieved. (ab: abnormal, norm: normal)

9. Exang (Exercise induced angina) Exercise induced angina (1 = yes; 0 = no).

10.Slope (The slope of ST segment) The slope of the peak exercise ST segment. Value 1: upsloping. Value 2: flat. Value 3: down sloping.

11. CA (Number of vessels coloured) Number of major vessels (0-3) coloured by fluoroscopy.

12. Thal (Obtained defect) 3 = normal; 6 = fixed defect; 7 = reversible defect.

13. Old peak value ranging from 0 to 6.

14. Class is either healthy (0) or with heart-disease (1).

**Algorithms Implemented**

Out of the various supervised learning algorithms, we used some algorithms which are suitable.

1. KNeighborsClassifier

2. Support Vector Machine

3. DecisionTreeClassifier

4. Random Forest Classifier

5. Voting Classifier

The general steps that are followed in every algorithm are:

1.Set up training data.

2.Set up the testing data.

3.Set up model parameters.

4.Train the model.

5.Predict the test data.

6.Determining accuracy and classification report.

Considering the given dataset. We have split the dataset into 2 parts i.e. Training Set and Testing Set. We can split it into any fraction. The best suitable one is 75% Training Set and 25% Testing Set and further proceeded to next steps.

**K Neighbors Classifier:**

K nearest neighbors is one of the well-known algorithms for classification of data in machine learning.  
This does not presume any distribution of data whereas other classifiers assume distribution of data. such as gaussian mixture model presumes data in the form of gaussian distribution. This non assuming behaviour of this algorithm makes it widely used in real-life situations this algorithm behaves as if the data is plotted on the graph and then divides it into clusters or groups. When a new unclassified data point is given it classify it based on which cluster or the group it is nearer. We chose neighbors either by brute force or by KD Tree.

In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

To select the K that’s right for your data, we run the KNN algorithm several times with different values of K and choose the K that reduces the number of errors we encounter while maintaining the algorithm’s ability to accurately make predictions when it’s given data it hasn’t seen before.

As we decrease the value of K to 1, our predictions become less stable. Inversely, as we increase the value of K, our predictions become more stable due to majority voting / averaging, and thus, more likely to make more accurate predictions (up to a certain point). In cases where we are taking a majority vote (e.g. picking the mode in a classification problem) among labels

In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors.

[12]

**The accuracy that we get using this model: 87%**

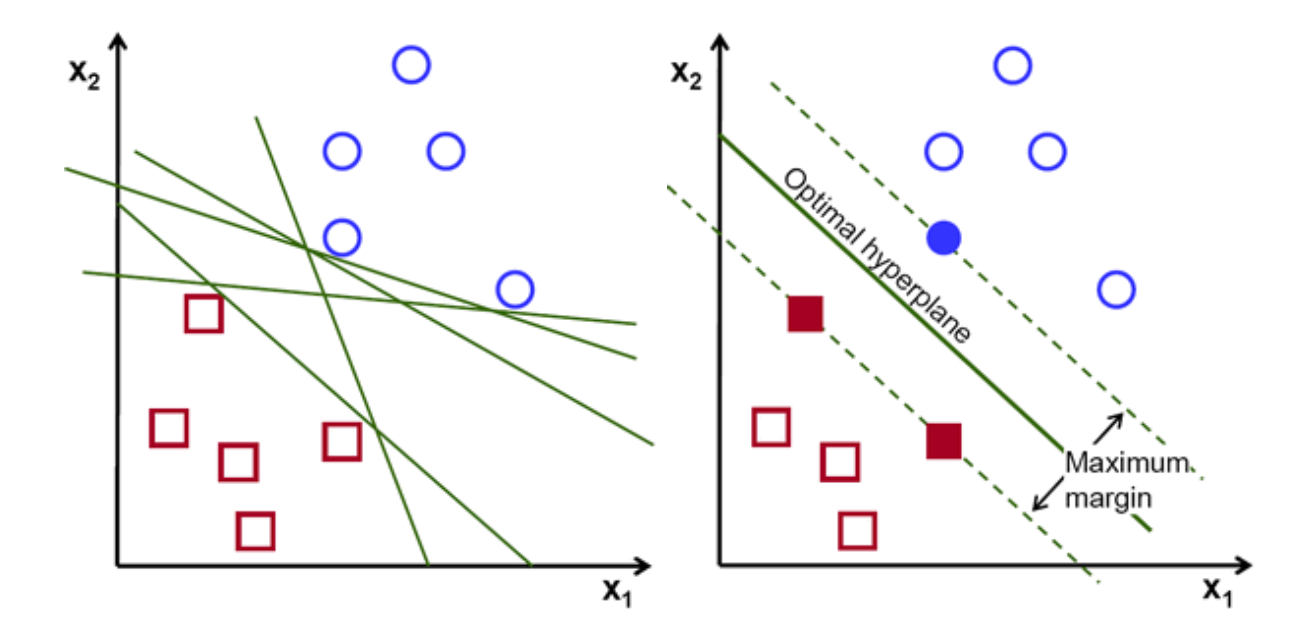
The accuracy that we got after using this model to train our model and predict for unknown data is 87%. It is reasonably good as this algorithm is very simple to implement and as we have less training data set it takes less time for computing.

**Support Vector Machine:**

“A SVM preforms classification by finding the hyper plane that maximise the margin between two classes. The vectors that define the hyper plane are the support vectors”.

To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features is 3, then the hyperplane becomes a two-dimensional plane. It becomes difficult to imagine when the number of features exceeds 3.

Given an input X and labelled target Y the algorithm outputs an hyperplane which then further classifies the new inputs.



[5]

There may be different types of margin to deal with. If there is linear boundary then it is linear support vector machine. There are other types of margin(hyper planes) like polynomial, gaussian etc.

Kernel, Regularization, Gamma and Margin are tuning parameters which then decide the algorithm working and accuracy according to the given dataset.

Kernels: Linear, rbf, sigmoid, polynomial

Regularization: The C parameter is the regularization parameter. For higher values of C the optimization will choose smaller margin.

**The accuracy that we get using this model: 84%.**

The accuracy that we got after using this model to train our model and predict for unknown data is 84%. It is very good as this algorithm is very powerful.

**Decision Tree Classifier:**

“A decision tree is a tree where each node represents a feature(attribute), each link(branch) represents a decision(rule) and each leaf represents an outcome (categorical or continues value)”.

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, and utility. It is one form to show an algorithm that simple contains conditional control statements.

A decision tree looks like a flowchart-like structure in which each internal node represents a “test” or “condition” on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the result of the test we then explore the branch in which the output falls in, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

In Decision Tree the major challenge is to identification of the attribute for the root node in each level. This process is known as attribute selection. We have two popular attribute selection measures:

1. Information Gain
2. Gini Index [6]

Information Gain:

When we use a node in a decision tree to partition the training instances into smaller subsets the entropy changes. Information gain is a measure of this change in entropy.



Entropy is the measure of uncertainty of a random variable, it characterizes the impurity of an arbitrary collection of examples. The higher the entropy more the information content.

Gini Index:

Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified. It means an attribute with lower Gini index should be preferred.

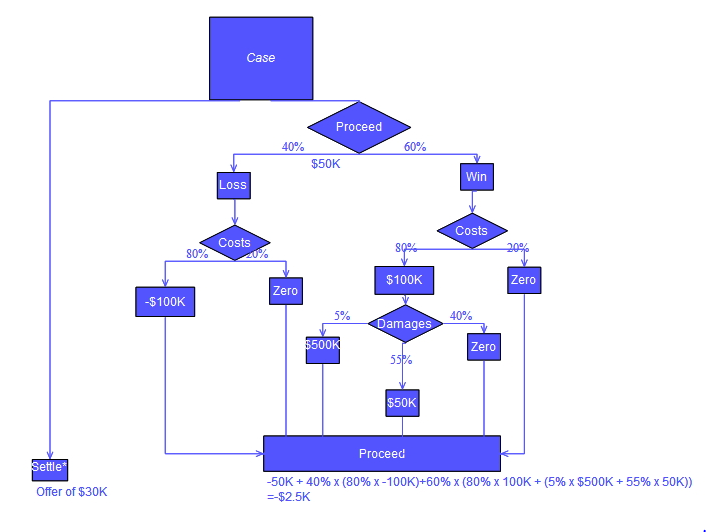
The Formula for the calculation of the of the Gini Index is given below.



Methods for generation of decision tree:

1. CART (Classification and Regression Trees): this method uses Gini Index (Classification) as parameter for classification of decision tree.

2. ID3 (Iterative Dichotomise 3): This method uses Entropy function and Information gain as parameter for building decision tree.



**The accuracy we got after using this model is : 84%**

The accuracy that we got after using this model to train our model and predict for unknown data is 84%.

Prone to overfitting.

Require some kind of measurement as to how well they are doing.

Need to be careful with parameter tuning.

Can create biased learned trees if some classes dominate.

**Random Forest Classifier:**

Random Forest Classifier is the extension of the Decision tree.

“Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.”

Some of the parameters for this classifier is number trees to form and the parameters that are used in making decision tree. In most of the cases it is better than general Decision tree classifier as it is generating many different decision trees and choose the best one.

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction.

“A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.”

The low correlation between models is the key. Uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. The reason for this wonderful effect is that the trees protect each other from their individual errors (as long as they don’t constantly all err in the same direction). While some trees may be wrong, many other trees will be right, so as a group the trees are able to move in the correct direction. So the prerequisites for random forest to perform well are:

There needs to be some actual signal in our features so that models built using those features do better than random guessing.

The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

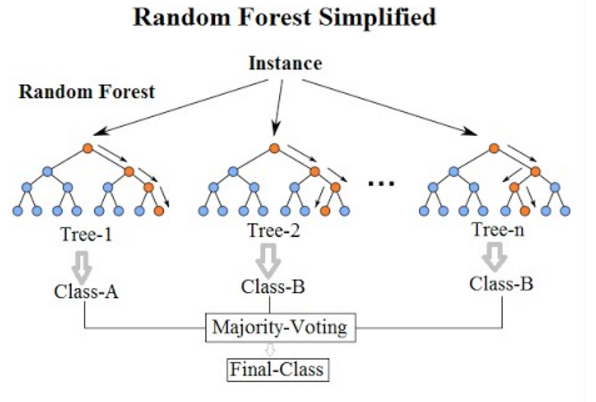
In a normal decision tree, when it is time to split a node, we consider every possible feature and pick the one that produces the most separation between the observations in the left node vs. those in the right node. In contrast, each tree in a random forest can pick only from a random subset of features. This forces even more variation amongst the trees in the model and ultimately results in lower correlation across trees and more diversification.

So, in our random forest, we end up with trees that are not only trained on different sets of data (thanks to bagging) but also use different features to make decisions.

Point needed in order to make accurate class predictions

We need features that have at least some predictive power. After all, if we put garbage in then we will get garbage out.

The trees of the forest and more importantly their predictions need to be uncorrelated (or at least have low correlations with each other). While the algorithm itself via feature randomness tries to engineer these low correlations for us, the features we select and the hyper-parameters we choose will impact the ultimate correlations as well.



[7]

**The accuracy we got after using this model is: 84%**

The accuracy that we got after using this model to train our model and predict for unknown data is 84%. It is more powerful when compare to decision trees

But our accuracy does not vary when compared to the normal decision tree as our does not vary much and has limited number of record.

**Voting Classifier:**

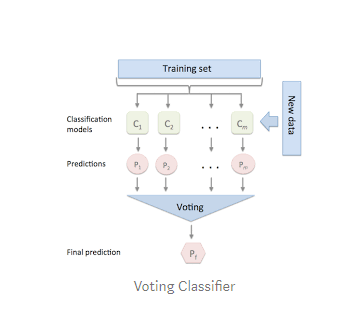
A collection of several models working together on a single set is called an ensemble. The method is called Ensemble Learning.

Voting is one of the simplest ways of combining the predictions from multiple machine learning algorithms. Voting classifier isn’t an actual classifier but a wrapper for set of different ones that are trained and valuated in parallel in order to exploit the different peculiarities of each algorithm.

We can train data set using different algorithms and ensemble then to predict the final output. The final output on a prediction is taken by majority vote according to two different strategies:

**Hard voting / Majority voting**: Hard voting is the simplest case of majority voting. In this case, the class that received the highest number of votes Nc​(y​​t)will be chosen. Here we predict the class label y^via majority voting of each classifier.

**Soft voting**: In this case, the probability vector for each predicted class (for all classifiers) are summed up &averaged. The winning class is the one corresponding to the highest value

[8]

**The accuracy we got after using this model is: 87%**

The accuracy that we got after using this technique on our data it gave 87%. As it takes generalizes the decisions taken by many algorithms it has

1. Lower error
2. Less Over fitting
3. Taste great

**Results**

K Nearest Neighbors:

Without cross validation

|  |  |
| --- | --- |
| **K value** | **Accuracy** |
| 1 | 0.76 |
| 2 | 0.78 |
| 4 | 0.84 |
| 6 | 0.86 |
| 8 | 0.87 |
| 10 | 0.83 |

With cross validation for the best k value k=8: **79.82%**

Support Vector Machine:

Without cross validation

|  |  |
| --- | --- |
| **K value** | **Accuracy** |
| 1 | 0.76 |
| 2 | 0.78 |
| 4 | 0.84 |
| 6 | 0.86 |
| 8 | 0.87 |
| 10 | 0.83 |

With cross validation for best of kernel = linear: **80.450%**

Decision Tree:

Without Cross Validation

|  |  |
| --- | --- |
| **Max Features** | **Accuracy** |
| 1 | 0.75 |
| 4 | 0.71 |
| 7 | 0.81 |
| 11 | 0.76 |
| 14 | 0.68 |

With cross Validation for the best resulted max features: **69.8%**

Random forest:

Without Cross Validation

|  |  |
| --- | --- |
| **Estimators** | **Accuracy** |
| 10 | 0.79 |
| 50 | 0.81 |
| 100 | 0.84 |
| 300 | 0.83 |

With cross Validation for the best resulted max: **78.156%**

Voting Classifier:

Without cross Validation for the best resulted max features: **87%**

With cross Validation for the best resulted max features: **80%**

**Future Scope**

When the models are trained, we will test the model with the testing data and calculate accuracy and precision. We will choose the model which gives higher performance in classifying the new data.

From the above accuracy results we can say that “Voting Classifier” relatively gave the better performance among the other algorithm used.

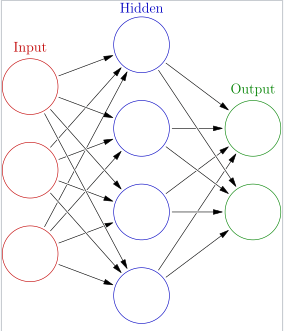
Now when we get a new patient with all the above-mentioned attributes or data, we can predict the occurrence of heart disease based on the outcome of the model. We can take subsequent precaution steps in saving the life of patient if there is change of occurrence of heart disease near future.

If we could collect the data of more patient, then we can more accurately predict the outcome. The future scope of this analysis would be helpful for the detection of heart disease for a patient

We can try one of deep learning algorithm neural network and can use it if it has more accuracy and precision than the above algorithms used. It requires a larger dataset to train well and predict well and get more accuracy.

**Artificial Neural Networks:**

“The idea of Artificial neural networks (ANN) is taken from the concept of human brain or neural network of the human. It tries to replicate the behaviour of biological neural networks. But they are not the same as biological neural networks. This ANN systems "learn" to perform tasks that is given by observing the examples, generally without being programmed with task-specific rules”.



ANNs began as an attempt to exploit the architecture of the human brain to perform tasks that conventional algorithms had had little success. They soon reoriented towards improving empirical results, mostly abandoning attempts to remain true to their biological precursors. Neurons are connected to each other in various patterns, to allow the output of some neurons to become the input of others. The network forms a directed, weighted graph.

**References**

1. Heart Disease Prediction Using Effective Machine Learning Techniques Avinash Golande, Pavan Kumar T.
2. Heart Disease Prediction using Logistic Regression Algorithm using Machine Learning
3. <https://en.wikipedia.org/wiki/Supervised_learning>
4. <https://en.wikipedia.org/wiki/Unsupervised_learning>
5. <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>
6. <https://towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96>
7. <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>
8. <https://medium.com/@sanchitamangale12/voting-classifier-1be10db6d7a5>
9. <http://www.iraj.in/journal/journal_file/journal_pdf/6-71-140490825388-92.pdf>
10. <https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72>
11. <https://medium.com/deep-math-machine-learning-ai/chapter-4-decision-trees-algorithms-b93975f7a1f1>
12. <https://medium.com/@srishtisawla/k-nearest-neighbors-f77f6ee6b7f5>